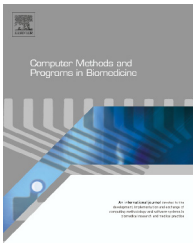
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 126 (2016) 128–142

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk

P. Fuster-Parra[a,b,*], *P. Tauler*[b], *M. Bennasar-Veny*[b], *A. Ligęza*[c], *A.A. López-González*[d], *A. Aguiló*[b]

[a] *Department of Mathematics and Computer Science, Universitat Illes Balears, Palma de Mallorca, Baleares E-07122, Spain*

[b] *Research Group on Evidence, Lifestyles & Health, Research Institute on Health Sciences (IUNICS), Universitat Illes Balears, Palma de Mallorca, Baleares E-07122, Spain*

[c] *Department of Applied Computer Science, AGH University of Science and Technology, Kraków PL-30-059, Poland*

[d] *Prevention of Occupational Risks in Health Services, GESMA, Balearic Islands Health Service, Hospital de Manacor, Manacor, Baleares E-07500, Spain*

ARTICLE INFO

Article history:
Received 18 August 2015
Received in revised form
28 November 2015
Accepted 11 December 2015

Keywords:
Bayesian networks
Model averaging
Cardiovascular lost years
Cardiovascular risk score
Metabolic syndrome
Causal dependency discovery

ABSTRACT

An extensive, in-depth study of cardiovascular risk factors (CVRF) seems to be of crucial importance in the research of cardiovascular disease (CVD) in order to prevent (or reduce) the chance of developing or dying from CVD. The main focus of data analysis is on the use of models able to discover and understand the relationships between different CVRF. In this paper a report on applying Bayesian network (BN) modeling to discover the relationships among thirteen relevant epidemiological features of heart age domain in order to analyze *cardiovascular lost years* (CVLY), *cardiovascular risk score* (CVRS), and *metabolic syndrome* (MetS) is presented. Furthermore, the induced BN was used to make inference taking into account three reasoning patterns: *causal* reasoning, *evidential* reasoning, and *intercausal* reasoning. Application of BN tools has led to discovery of several direct and indirect relationships between different CVRF. The BN analysis showed several interesting results, among them: CVLY was highly influenced by smoking being the group of men the one with highest risk in CVLY; MetS was highly influence by physical activity (PA) being again the group of men the one with highest risk in MetS, and smoking did not show any influence. BNs produce an intuitive, transparent, graphical representation of the relationships between different CVRF. The ability of BNs to predict new scenarios when hypothetical information is introduced makes BN modeling an Artificial Intelligence (AI) tool of special interest in epidemiological studies. As CVD is multifactorial the use of BNs seems to be an adequate modeling tool.

* *Corresponding author at*: Department of Mathematics and Computer Science, Universitat Illes Balears, Palma de Mallorca, Baleares E-07122, Spain. Tel.: +34 971171386.
E-mail address: pilar.fuster@uib.es (P. Fuster-Parra).

http://dx.doi.org/10.1016/j.cmpb.2015.12.010
0169-2607/

1. Introduction

Bayesian Networks (BNs) [1,2] also referred to as *Belief Networks* or probabilistic causal networks are an established framework for uncertainty management in Artificial Intelligence (AI). They constitute a tool which combines graph theory and probability theory to represent relationships between variables (nodes in the graph) [3]. Contrary to deterministic understanding of the *causality* phenomenon [4], BN modeling has its origins within data mining and machine learning research [5,6] and captures probabilistic influences induced out of big data sets. They constitute a powerful knowledge representation and an efficient reasoning tool under conditions of uncertainty [7]. The network structure is a directed acyclic graph (DAG) where each node represents a random variable [8,9] and the arcs are suitable for representing causality [10].

BNs have been proven to be a strong tool to discover the relationships between variables that attempts to separate out direct and indirect dependencies [11,12], and can capture the way an expert understands the relationships among all the features [13]. BN modeling is widely used in fields like clinical decision support [14], systems biology [15,16], human immunodeficiency virus (HIV) and influenza research [17,18], analyzes of complex disease systems [19–21], interactions between multiple diseases [22], and also in diagnostic diseases [23–27].

The metabolic syndrome is a set of risk factors that include abdominal obesity, insulin resistance, dyslipidemia and hypertension leading to increased risk of developing cardiovascular diseases and type 2 diabetes [28–31]. Cardiovascular disease (CVD) epidemiology is a worldwide public health problem [32]. The economic burden of CVD is already affecting the economies of the world's wealthiest countries. However, in the next decades developing countries will be more affected due to the great increase in CVD prevalence expected in these countries [33]. It is estimated that in 2015, more than 20 million people may die worldwide because of CVD. This number is expected to increase in the upcoming decades, that every 5 s in the world a myocardial infarction would occur [34,35].

CVDs are closely related to the well-known cardiovascular risk factors (CVRF). The concept of CVRF appeared in 1961, when the group of Kanned defined CVRF as biological traits or behaviors that increased the chance of developing or dying from CVD [36,37]. The high prevalence of certain risk factors to which we are exposed is the cause of this situation, in which the prevalence of CVD is increased every year. It is necessary to control the factors that influence the development of CVD, such as smoking, hyperlipidemia, hypertension, diabetes, obesity, a diet high in saturated fats, alcohol abuse, a sedentary lifestyle, and stress [38]. In fact, WHO (World Health Organization) estimates that 80% of premature deaths from cardiovascular disease and diabetes could be prevented by efficient controlling these risk factors [39].

There are some scores that numerically quantify cardiovascular risk (CVR). One of the most widely used is Framingham score, with its calibrated form for the Spanish population, the Framingham-REGICOR [35]. This scale estimates the global CVR to 10 years and it is expressed as a percentage. Recently, a new score has been proposed, the so-called Heart Age tool (HA), which is based on Framingham score, and supposes a simple and graphic way to communicate the CVR because it expresses the CVR as an age. If the HA value is older than chronological age the term "lost years", defined as the HA minus the chronological age, could be used. The HA is a novel concept designed specifically to help people to understand their own cardiovascular disease risk and implement changes into their lifestyles to prevent the incidence of CVD [40].

Development and analysis of models to examine the relationships between different CVRF could be not only of theoretical interest, but can serve as a generic tool for application oriented activities: explanation, prediction, monitoring and prevention. It enables both theoretical analysis of the relationships between numerous variables, and having in mind the probabilistic nature of the causal dependencies, BNs seem to be an adequate tool. Moreover, BN models are capable of creating different scenarios based on hypothetical cases when new observations are instantiated.

The paper is organized as follows. Section 2 introduces BNs and some basic concepts for inference flow. Section 3 presents the materials and methods for the epidemiologic study and the process of inducing a BN from a data set. Section 4 shows different reasoning patterns to analyze the BN. Section 5 presents a discussion. Finally, Section 6 concludes the paper.

2. Bayesian networks

A BN consists of [41]: (i) a set of variables and a set of directed edges between these variables, where (ii) each variable has a finite set of mutually exclusive states, and (iii) the variables together with the directed edges form a DAG. BN models estimate the joint probability distribution P over a vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)$. The joint probability distribution factorized as a product of several conditional distributions denotes the dependency/independency structure by a DAG:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i^{\mathcal{G}})) \tag{1}$$

Eq. (1) (where $Pa(X_i^{\mathcal{G}})$ denotes the parent nodes of X_i) is the main reason for the formulation of a multivariate distribution by BNs; this equation is also called the *chain rule for Bayesian networks*.

As BNs are used to make inference [8], it is necessary to understand the flow of influence when new information is introduced in a BN. Below we introduce some basic concepts.

Two variables X and Y in a BN are *d-separated* if, for every possible path between X and Y, there is an intermediate variable Z such that either: (i) the connection is serial ($X \to Z \to Y$ or $X \leftarrow Z \leftarrow Y$) or diverging ($X \leftarrow Z \to Y$) and Z is instantiated, or (ii) the connection is converging ($X \to Z \leftarrow Y$) and neither Z nor any of Z's descendants have received evidence. When influence flows from a node X to another node Y via a node Z, it is said that the trail $X \rightleftharpoons Z \rightleftharpoons Y$ is active. A causal trail $X \to Z \to Y$ (serial connection), an evidential trail $X \leftarrow Z \leftarrow Y$ (serial connection) or, a common cause trail $X \leftarrow Z \to Y$ (diverging connection) is active if and only if Z is not observed. A common effect trail